Judging Research Papers for Research Excellence

Journal:	*Studies in Higher Education*
Manuscript ID	CSHE-2016-0295.R1
Manuscript Type:	Article
Keywords:	research accountability, Rasch, reliability, academic articles, Research Excellence Framework

Judging Research Papers for Research Excellence

Abstract

The UK's Research Excellence Framework of 2014 was an expensive high stakes evaluation which had a range of impacts on higher education. One component was an assessment of the quality of research where a series of panels read and rated the outputs of their peers. Quality control was strengthened after the exercise of 2008, but questions still remain about how fair it is to rate all papers on the same scale by raters who may vary in their reliability and severity.

This paper takes data from a large department in which 23 senior staff rated the outputs from 42 academics. The analyses, using the Rasch model, showed that: a single scale described the data well; most raters were reliable; there was a noticeable variation in the severity of the raters.

Suggestions for future exercises include a pre-appointment procedure for panel members and statistical adjustments for the severity/leniency of raters.

Keywords: word; research accountability, Rasch, reliability, academic articles

Introduction

The Research Excellence Framework (REF) is the system for assessing the quality of research in UK higher education institutions (HEIs) (http://www.ref.ac.uk/). As in previous Research Assessment Exercises (RAE) in 1996 and 2008, secondary peer review by domain specific panels (Units of Assessment) played an important role in evaluating the quality of written 'outputs' (which were mainly peer-reviewed journal articles but also books, chapters in edited books and research reports: see Bence & Oppenheim, 2004). The assessment outcomes inform the selective allocation of £1.6 billion public funding for research to the HEIs from 2015 until whenever the next assessment is undertaken or an alternative is introduced. Additionally, the aim of the assessment exercise is to provide accountability for public investment in research and to produce evidence of the benefits of this investment. A study commissioned by the

Higher Education Funding Council for England (HEFCE) in 2008 estimated the RAE to cost HEIs in England approximately £47 million (PA Consulting, 2009). The REF Accountability Review calculated the total cost of REF2014 at £246m, with £14m for funding bodies (HEFCE and its counterparts for Scotland, Wales, and Northern Ireland) and £232m for HEIs. Around £19m of this was identified for REF panellists' time as peer-reviewers.

Peer review is a questionable approach in terms of its reliability (Newton, 2010). Cicchetti (1991) found in overall assessments based on 16 peer-review studies of journal articles that individual raters' reliabilities varied from 0.19 to 0.54 (median 0.30). Marsh and Ball (1989) found that individual reviewers' reliability for the overall recommendation in the Journal of Educational Psychology was 0.30 between reviewers compared with an average of 0.27 found in other research they reviewed. They also indicated that an individual's idiosyncratic response biases (such as a tendency toward leniency or harshness in their evaluation) were larger for the specific sub-scale rating items than for the overall recommendation. Marsh et al. (2008) found that peer-review of grant proposals was similarly unreliable, requiring at least six raters to reach a reliability of 0.70, though this could be improved by using a smaller number of raters who had similar expertise relevant to the content of the proposal. It has also been noted that peer-review tends to be conservative and promotes the status quo in terms of assessments of quality and allocation of grants and resources (Squazzoni & Gandelli, 2012; Hamann, 2016).

The REF assessment of research quality in the UK is mirrored in many other countries such as the Netherlands, France and Australia (see Rebora & Turri, 2013 for an interesting comparison of the UK's RAE and Italian VQR (Valutazione della qualitàdella ricerca or Research Quality Assessment)). REF 2014 was the seventh in

the series of government-mandated evaluations of research quality in each department (or at least across defined disciplinary 'units of assessment') in every UK University. The objective has been to manage the allocation of central research funding to 160 universities. It was called the 'Research Assessment Exercise' (RAE) on six previous occasions, and Hicks (2009) estimated that the outcomes determined approximately 25% of all research support in UK universities. Overall, the funding allocations have remained stable. After the 2001 RAE, only one institution's income varied by more than 3.7 per cent and the median impact was less than 0.6 per cent (Sastry & Bekhradnia, 2006).

The RAE and REF methodology has evolved and developed into an increasingly complex and expensive process (Stuart, 2014). At its core there remains a peer evaluation of departmental research 'outputs'. In 1986, Departments or 'Units of Assessment' (UoA) described their research achievements in two pages, listed their five best publications and provided data on research income, and indicators of esteem. This approach was criticized for favouring larger departments able to select the top five papers from a larger pool. In response to these kinds of criticisms, the approach has evolved so that submissions now include greater detail on research environment, strategy and impact with up to four publications or 'outputs' for each individual included and other data on research income and research student completions. For a small department the required commentary to accompany the data now runs to about 24 pages of text.

How the REF assessments were made

A useful account and overview of how quality was assured for REF 2014 is given in *Research Excellence Framework 2014: The results* (REF 2014). It set out the

rules, criteria and working methods. There is also a set of Overview Reports (see for example Overview Report by Main Panel C: REF 2015).

To provide an illustration of the work of a sub-panel, we were able to consult Professor Andrew Pollard who chaired the Education Sub-panel for REF 2014 on behalf of UK HEFCEs. Whilst each panel may have operated in different ways, it is thought that their general approaches were similar. Pollard explained the importance of three major factors – a managed sequence of interactive processes associated with peer review for the initial calibration and final moderation of judgements, the work of individual panellists in exercising academic judgement, and a system for numerical monitoring of assessment progress and proposed assessment grades.

A key feature of the quality assurance was the sheer number of meetings that were coordinated. For Main Panel C, 11 sub-panels worked together to share the current position and analyses of their data. They met on eight full-day meetings spread over a 10 month period. Interspersed with these, the sub-panels met seven times, with a mixture of meetings lasting one, two or three days – in the case of SP25, Education, meeting for 13 days in total. There was thus regular iteration between the main and sub-panels as they progressed through a coordinated sequence of output calibration exercises, assessment activity, moderation and confirmation of grades and profiles. Within the Education sub-panel there was a great deal of interaction between panellists both at face-to-face meetings (plenary, group or paired) and through the confidential email system which had been made available by the funding councils and REF managers.

In relation to the assessment of outputs every paper was rated on a five point scale from unclassified through 1* (recognised nationally), 2* (recognised internationally), 3* (internationally excellent) to 4* (world leading). An initial training

and calibration process took place between the sub-panel chairs within each Main Panel. This was then replicated within each sub-panel. Thus members of the Education sub-panel spent seven weeks with all members commenting on and assessing 'outputs of the week', topped off with a two day face-to-face meeting to discuss agreements and differences in proposed grades. In the main assessment phase, some outputs were assessed by several panellists, others by just two and a significant proportion by just one. Confidence in the initial calibration, and later, in moderation processes, was thus very important.

Throughout the assessment process, a system of linked Excel spreadsheets provided numerical information on the progress of assessment and on patterns in the evolving judgements. The Excel system enabled management of the entire REF operation and was based on multiple, nested worksheets which, when uploaded and downloaded in sequence, integrated and refreshed both the collation and analysis of workload and assessment data. Administrators or managers were at the top level and below them were the four main panels, then the thirty-six sub-panels and then the more than 1,000 members, assessors and specialist advisors. Each individual at the lowest level had a worksheet which they downloaded and into which they entered their ratings before uploading them. Access to the nested spreadsheets was available according to position in the four-level hierarchy described above. The chair of a sub-panel could see all data on that sub-panel but not the main panel data.

A statistical procedure was applied automatically to the data so that panel chairs could see the percentage of submissions that had been rated, broken down by books and papers, and could see the means and the standard deviations by panellist and groups within the UoA. If a chair felt that something wasn't quite right such as particularly high or low averages or an unexpected standard deviation, he or she was able to initiate

a discussion, send an email and/or propose a re-assessment. The Education sub-panel Chair essentially saw himself as orchestrating a process which drew on, and honoured, the particular expertise of panellists but was also watchful for any patterns in proposed grades which might not be justified. Where there was apparent leniency or severity by individual raters, he would make them aware of this and invite consideration of the issues. This took place routinely by email, and, if judged appropriate, was also followed up through the appointment of additional assessors, or in group or plenary discussion at face-to-face meetings. In due course, the pattern of outcomes from the sub-panels was monitored by the Main Panels and similar moderation processes were adopted. In particular, there was regular review of the means and the standard deviations across the sub-panels, with justifications sought where they were felt to be necessary. Thus in all settings, whether in the sub-panels or Main Panel, variation in patterns between sub-fields or assessors was noted and investigated – with the most notable examples being the subject of plenary discussion. Where patterns of variation were felt to be academically justified, they remained, but where variation was not supported, ultimately, a plenary judgement was obtained. Indeed, the REF system was explicit that the recommended final grades and overall quality profiles were the responsibility of the sub-panel members as a whole. Members collectively considered and approved final outcomes.

For the 2008 RAE, some strategies had been used to achieve comparability between sub-panels but evaluations had suggested that there was scope for improvement. For REF2014 there were thus very deliberate strategies to develop comparability through various forms of co-ordination across the exercise. The use of the Excel spreadsheet system to analyse and compare results provided an important backbone to the exercise.

The context for the collection of data

The analysis presented in this paper is of data from the School of Education at Durham University, where staff involved in the review of papers also included staff from the Centre for Evaluation and Monitoring (CEM) and the Centre for Medical Education Research, as eligible for submission in the Education Unit of Assessment in the REF. Faced with a possible decision at University level to return a limited number of staff, it was important to get an accurate picture of papers for possible inclusion. Great efforts were made to this end and this process generated some useful data to help with a decision-making process which might affect individual staff significantly as well as support judgements about the development of the Departmental REF submission.

Research Questions

A number of key questions underpinned the analysis. In particular it was important to identify the extent of agreement between raters, the consistency of raters and the leniency/severity of the raters compared to one another. A parallel series of questions relates to the ratings of the papers in terms of how many categories the raters can distinguish reliably, and whether some papers produced more agreement from raters than others. This therefore included understanding whether some raters appeared to be assessing different aspects of the papers (i.e. Was there evidence for more than one construct?). A third series of questions related to the authors of the papers in terms of how consistent the ratings of their papers were. To what extent do the papers from individuals hang together? Then finally the key questions were: What were the implications for a department preparing a submission and what were the implications for the workings of the REF panels?

Ethical approval

The data were collected as part of the Department's preparations for REF 2014. Ethical approval for use of the data to undertake the analysis presented in this paper was granted by the School of Education's Ethics Committee at Durham University. Inclusion in this analysis was by opt-out, with removal of the data of anyone who requested not to be included. One person out of 43 requested not to be included so their data were removed from the analysis.

Chosen approach to the analysis

The analysis of the data was challenging in a number of ways. Some academics were likely to have seen higher quality papers than others and they were likely to vary in the severity of their ratings. There was some overlap in the papers which the academics rated and the analysis would, ideally, use these overlaps to make adjustments for the severity of ratings by the academics and, possibly, exclude unreliable raters. This should lead to better estimates of the quality of papers, and the severity/leniency of the raters. The analysis should also test the assumption, implicit in REF, that a single scale running from unclassified to 4* is justified. Fortunately, there is a statistical modelling technique which is uniquely suited to these challenges. It involves putting rater severities and paper quality onto the same scale. It does this iteratively and, in doing so, it provides tests for how well raters and papers fit the model of a single scale. In other words we can readily seek answers to the questions set out above.

This technique is the Rasch model. It is widely used for the development of assessments and the analyses of assessment-based data because of the way that it manages the trade-off, typically between a test-taker's performance and the item difficulties. In this paper the trade-off is the quality of the paper submitted for the REF,

in relation to the severity or leniency of the rater. Approaches such as inter-rater reliability may identify some differences, but will not be able to scale the scoring appropriately. From a slow start after George Rasch carried out his seminal work in the 50's (Rasch, 1993) there has been an exponential growth in published papers which refer to Rasch measurement (Panayides et al., 2015). A valuable introduction is provided by Bond and Fox (2015).

The data collection

All staff were asked to nominate up to six research papers or 'REF outputs' in accordance with the REF criteria, although some individuals submitted more than this number. Excluding the one academic who opted out of the exercise, there were data on 42 academics who had authored 223 papers (1 to 223). Their papers were assessed by 23 raters (A to W). Table 1 below gives some details of the numbers of papers per academic.

Insert Table 1 about here

All of the professorial staff and Management Committee were asked to be raters as well as four external raters selected to provide feedback on the Department's developing REF submission by the University. A further invitation was sent to all the academic staff in the School of Education to which three staff members responded. The external reviewers, who included two former RAE panel members, rated a 10% sample of papers. Two of the departmental reviewers were also later involved in the REF 2014 analysis of papers.

Each person was asked to rate a range of papers with the aim that every paper would be reviewed by at least two raters, though this was not always possible and some were rated by more than this, see Table 2 below. So far as possible the raters were given a range of papers to assess but given the tight schedule in the run up to the REF2014 deadline it was not possible to randomly assign papers to raters.

Insert Table 2 about here

Preparation for raters

The raters were not given formal training rather they were given the criteria (http://www.ref.ac.uk/panels/assessmentcriteriaandleveldefinitions/) and asked to use those details to make their judgements. In line with the guidelines they were instructed not to be influenced by who had written the article nor by where it was published nor whether it was backed by a research grant. They were also asked not to discuss the papers that they were rating. All the raters had had experience of the previous RAE, all had experience of reviewing articles for academic journals and proposals for grants and all had examined PhDs.

The numbers of ratings given to each level are shown in Table 3.

Insert Table 3 about here

The analysis

An initial impression of the data is given in Figure 1 which shows the distribution of the average ratings of all 223 papers.

Insert Figure 1 about here

It shows that the full range from unclassified (0) to 4* was used and that the distribution was approximately normal although the majority lay from 1* to 4* inclusively. The average rating was 2.4 and the standard deviation was 0.71.

The parallel distribution for the ratings given by the 23 raters is shown in Figure 2.

Insert Figure 2 about here

Again the distribution was approximately normal with a mean of 2.4. The standard deviation was lower at 0.35.

The two figures (1 & 2) raise questions which suggest that they should not be interpreted directly. Did the wide range of ratings in Figure 2 have its origin in the papers that the raters rated or did the raters vary more systematically in the severity or leniency of their ratings, or both? Was the rating scale based on a single construct? These and other queries mentioned in the introduction are addressed below.

The software packages used for the analyses were Winsteps version 3.90.0.0 for Rasch measurement and MLwin version 2.35 for the multilevel modelling.

How many categories?

Because the raters were not restricted to whole categories and many gave fractional ratings (such as 2/3), it was necessary to decide if fractional ratings should be used. An exploratory analysis using the Rasch rating scale model included all data and the same scale for all raters. It indicated that the categories used by the raters were not separable from one another at any finer level than whole numbers i.e. 0, 1*, 2*, 3* and 4*. The chart below (Figure 3) shows the issue visually for all raters. On the left the probability of a paper being given each point of a nine point scale corresponding to the inclusion of half grades is shown. It is clear that the half points were never the most

probable outcome and the data thereafter were analysed having rounded off the fractional grades (halves were rounded down). The chart on the right shows that now each of the points 0 to 4* were the most likely rating for some level of the estimated paper measure.

Insert Figure 3 about here

Reliability and fit

Including all papers and raters in the Rasch model produced reliability estimates as follows:

Number of papers measured = 224: Paper reliability 0.76

Number of raters measured = 23: Rater reliability 0.87

Two raters out of 23 did not fit the model well as indicated by the fit statistics. A high Outfit Mean-square figure (>2) and low point-biserial correlation (<.5) were used to make this judgement). This misfit can be illustrated by considering the ratings of paper 54. Rater L, who was identified as misfitting, gave the paper a 2* but raters B and O gave it a 4*. Of course, this could be due to severity of Rater L but this lack of agreement was not consistent and is summarised by the correlation of L's ratings with the overall Rasch estimated measure. It was 0.36; a low figure. The other misfitting rater (J) had a correlation of 0.44. The rest of the raters had a mean correlation of 0.82 (SD 0.10) indicating a high level of agreement between them.

Similarly 35 papers out of 223 did not fit the model well. One example is provided by paper 16 which was given ratings of 2,4,0,2 by raters E, G, K and W who were considered to be reliable.

Despite these misfitting papers and raters the model indicated a considerable degree of agreement. For the papers assessed by two or more raters 25% were

unanimous in their ratings and for 72% of papers the ratings did not differ by more than one point on the REF scale.

How do the Rasch measures compare with unadjusted ratings?

The Rasch measures of papers correlated with the average raw rating of each paper at the 0.97 level, which is very high, and the scattergram in Figure 4 confirms a near linear relationship.

Insert Figure 4 About here

Despite the strong agreement there are some clear differences. For example, of the papers which were given a logit of score -2 on the Rasch measure, one had an average rating of 2* and the other of 1*.

Errors of measurements

Discrepancies may be partially explained by errors of measurement and a nice feature of Rasch is that it estimates the error of measurement for each paper. The error depends strongly on the number of ratings which a paper is given and it also depends on the level. This is shown in Figure 5. There is also the possibility that there are links between the raters' own academic level and the papers that were rated but, for confidentiality reasons, we were unable to test that.

Insert Figure 5 about here

The chart indicates that papers rated four or more times had a standard error of between 0.18 and 0.3 of a level. On the other hand if only one person rated the paper then the standard error is likely to above 0.5.

The chart also shows that weaker papers were more reliably measured (lower standard error); there was more agreement amongst raters at the lower end of the scale, a point noticed by Cicchetti (1991).

Dimensionality

A check was made to see if one construct was sufficient to describe the data. A principal component analysis of residuals for raters (Wright, 2000) produced an eigenvalue of 1.9 for the first contrast indicating that a second dimension might be present; this is above the figure of 1.4 which the simulation work of Smith and Miao (1994) suggested as a cut-off, although it is recognised that each dataset is different. The contrast largely corresponded to raters A and B and suggested that they took a slightly different perspective from the rest of the raters. The second and third contrasts were also above 1.4. Overall the data were clearly dominated by a single construct. 72% of the variance was "explained" by the measures, and only 2.4% by the first and 2.2% by the second contrast.

Leniency and severity of ratings

There was clear evidence that the raters differed in their leniency/severity and this is shown in the map of papers and raters in Figure 6. This figure holds a considerable amount of information and essentially summarises the main results of the analysis.

Insert Figure 6 about here

How to read Figure 6.

The vertical line below the word MAP represents the equal interval log odds (logits) scale. Just to the left and right of the vertical line are the letters M, S and T

which mark the mean, one standard deviation and two standard deviations around the mean for papers and raters.

To the left, each paper is located by a # or a . at its measure. A # represents two papers and a . is one paper. To the right each rater, A to W, is located at his or her estimated measure. The raters higher up the scale are more severe and those lower down are more lenient. The horizontal arrows show the thresholds between adjacent outcomes. The lowest is for 0 (unclassified) to a 1*. Just two papers fall below that threshold. The highest is for the threshold between a 3* and 4*. Twelve papers appeared above the threshold. .

What Figure 6 shows

- The papers follow a similar distribution to the raw data shown in Figure 1 with most papers lying from 1* to 4* inclusively.
- The raters varied considerably in the severity/leniency of their ratings. In an ideal world these would all be equal but some raters were lenient and some severe in their assessments. The differences are large with M and G differing by about 7 logits on the Rasch measure. The former might rate a paper 1* whilst the latter might give the same paper a 3*
- The REF scale is ordinal but Rasch measurement creates an equal interval scale and the map indicates that the outcomes were not equally spaced. Cut scores are marked on the chart for the REF levels based on the Score Table from Winsteps. It is clear that the 1* covers a relatively small part of the scale (2.2 logits), the 2* covers a larger part (4.3 logits) and the 3* larger still (5.6 logits). It seems that it is easier to move from a 1* to a 2* than it is to move from a 2* to a 3*. The move from a 3* to a 4* is hardest of all.

Our interpretation here is that staff were encouraged to submit their best paper for the exercise, so the number of lower scoring papers may have been reduced. In addition it is hard to operationalise the criteria distinguishing 4* and 3*, suggesting raters were conservative in awarding the highest point on the scale.

Local dependency

The model included clusters of papers from the same academics and one would therefore expect there to be some local dependency. The question is: How much difference does this make to the patterns and conclusions that are reported? In order to answer this question Linacre's (2015) Procedure A was followed. The raw data were aggregated to give means at the academic level and where fractional rating appeared they were rounded off (halves were rounded down). The Rasch measurement results were then compared with the data produced using the paper as the unit of analysis. One consequence was that the spread of the scale in logits was reduced, but the correlation between the academic's measures and the average of the paper measures was 0.91 despite the loss of information due to rounding. The raters and academics generally fitted the model well although rater L again seemed to be assessing something different from the majority. The correlation of the raters' measures was also 0.91 between the two approaches. As expected different cut-points were found on the logit scale but the general patterns seen in paper based analysis and the academic based analysis were very similar.

How well are individuals separated out by their papers?

Having identified the best four papers for each academic using the Rasch measure, we asked how well the papers hung together and differentiated one academic from another? Using a multilevel model with papers nested within academics it was

estimated that 66.8% of the variance in paper measures was located with the academics. The variances at the paper level and the academic levels can be used to estimate the reliability of the measures of academics using the "shrinkage" formula from Goldstein (1987). This gives a figure of 0.89, if they had four papers. If they had three papers the figure dropped slightly to 0.86.

The average Rasch measure for each academic was then plotted in Figure 7 with an error calculated as the root mean square of the individual paper errors. This chart gives a visual indication of how well the data differentiated academics after the Rasch measures were converted to REF metric equivalents. It also provides a tool for decision making at the time of the REF submission. One point to note is that, whilst there are clear differences between high, middle and low rated academics, any one academic is indistinguishable, in any substantial sense, from his or her near neighbours. This implies that any cut-score for REF submission may be seen as unfair by an individual just below the cut score. This issue is particularly noticeable near the centre of the distribution.

Insert Figure 7 about here

Summary of the results of the analysis and reflections

In summary, the data generally fitted well into the Rasch model: there was one major dimension (construct) and just two raters out of 23 did not fit well with the view of others. Similarly the rating of 35 out of 223 papers did not fit the model.

The papers were measured reliably, but more accurately at the weaker end of the distribution and, as expected, more accurate estimates were made if there were more ratings.

Although the raters were generally reliable they varied markedly in the severity/leniency of their ratings.

Because anonymity is guaranteed we cannot display the characteristics of the raters but we were unable to find patterns which would explain reliability or severity. They appeared not to be related to academic standing, the holding of particular posts or depth of experience. Perhaps some raters tended to a literal interpretation of the level descriptors, (criterion referenced) and others were minded to see the exercise as comparative (norm referenced); this could explain some of the variation in severity. We are similarly unable to be explicit about why some papers did not fit the model, though it was clear that there were disagreements about whether some papers belonged to the UoA under consideration. Idiosyncratic responses may, of course, be an important part of the peer review process where 'herding' of judgements is a known phenomenon and may limit the development of scientific knowledge (Park et al., 2014) so any assessment of reliability also needs to consider other aspects of the validity of the judgements, particularly at the micro-level of peer-review (Oancea, 2007).

We also note that the data were very useful in talking with staff about their own papers and about their inclusion in the REF return. The data were also helpful in preparing the REF submission, such as by identifying papers which indicated a strong consensus about their quality. The external reviewers appointed by the University did not differ from the typical picture, which was interpreted as broad validation of the approach.

In terms of the results of the REF, we cannot tell how our ratings for individual papers matched the REF assessment. However the proportions in the submission for outputs were 26% at 4*, 45% at 3*, 27% at 2*, 2% at 1* (none were unclassified). This suggests that our cut off point for each category was slightly lower than anticipated (we

hoped for 30% 4*, 40% 3* and 20% 2*) with a larger proportion of 2* than we anticipated, but there were a larger number of papers clustered at this point on the scale. In addition the unit of submission is the individual, so a trade-off has to be made between volume and the range of outputs an individual might have available.

Implications

For departments and universities seeking to submit the best set of papers and academics, the methodology outlined in this paper might be useful. The one thing that we would add is that the exercise is probably best carried out on a continuous basis rather than as a one-off in the last year.

For future REF-successor panels there are more serious implications. Although the arrangements for REF were organised in such a way as to produce a common understanding of what levels meant within and across panels and sub-panels there remains the possibility that undetected severity or leniency remained especially in small units. Further mild leniency/severity may well have gone undetected or accepted even though it could have a major impact on league table positions. There was also no statistical mechanism to detect idiosyncratic judgements.

We propose that the selection of panel members could be informed by some mock assessments. Simply asking prospective members to rate half a dozen papers which are independently rated by six other experts in their area should give reassurance that reliable members are being appointed who are neither very severe nor very lenient. The second point is that, whilst we are confident that good statistical advice was available to the REF panel, future exercises may find it useful not to try too hard to ensure that all raters are of the same severity/leniency (as they never will be), but rather to accept the judgements, after extensive discussions and consultations, and to make nuanced statistical adjustments to take account of individual severity/leniency based on

sufficient data. The enemy of good assessment is not the leniency/severity of individuals which can be taken into account, but their unreliability.

References

Bence, V., & Oppenheim, C. (2004). The role of academic journal publications in the UK Research Assessment Exercise. *Learned Publishing*, 17(1), 53-68.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Routledge.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14, 119–135.

Goldstein, H. (1987) *Multi-level Models in Educational and Social Research.* London: Griffin School Effects.

Hamann, J. (2016/ online early) The visible hand of research performance assessment. *Higher Education*, 1-19.

Hicks, D. (2009). Evolving regimes of multi-university research evaluation. *Higher Education,* 57(4), 393-404.

Linacre, M. (2015) Winsteps software version 3.90.0.0 www.winsteps.com Help section

Marsh, H. W., & Ball, S. (1989). The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *The Journal of Experimental Education*, 57(2), 151-169.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *American Psychologist*, 63(3), 160.

Newton, D. P. (2010). Quality and peer review of research: an adjudicating role for editors. *Accountability in Research*, 17(3), 130-145.

Oancea, A. (2007). From Procrustes to Proteus: Trends and practices in the assessment of education research. *International Journal of Research & Method in Education*, 30(3), 243-269.

PA Consulting (2009) *RAE 2008 Accountability Review* London: PA Consulting Group

Park, I. U., Peacey, M. W., & Munafò, M. R. (2014). Modelling the effects of subjective and objective decision making in scientific peer review. *Nature*, 506(7486), 93-96.

Panayides, P., Robinson, C., & Tymms, P. (2015). Rasch measurement: a response to Goldstein. *British Educational Research Journal*, *41*(1), 180-182.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637.

Rebora, G., & Turri, M. (2013). The UK and Italian research assessment exercises face to face. *Research Policy*, *42*(9), 1657-1666.

REF (2014), *Research Excellence Framework 2014: The results* http://www.ref.ac.uk/media/ref/content/pub/REF%2001%202014%20-%20introduction.pdf downloaded 23/3/16.

REF (2015) *Research Excellence Framework 2014: Overview report by Main Panel C and Sub-panels 16 to 26* http://www.ref.ac.uk/media/ref/content/expanel/member/Main%20Panel%20C%20overview%20report.pdf downloaded 23/3/16.

Sastry, T. & Bekhradnia, B. (2006). *Using Metrics to Allocate Research Funds: A*

short evaluation of alternatives to the Research Assessment Exercise. (May, Oxford: Higher Education Policy Institute).

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.) *Objective measurement: Theory into practice. Vol. 2* (pp. 316-327). Norwood, NJ: Ablex.

Squazzoni, F., & Gandelli, C. (2012). Saint Matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics*, *6*(2), 265-275.

Stuart, D. (2015). Finding "good enough" metrics for the UK's research excellence framework. *Online Information Review*, 39(2), 265-269.

Wright, B. D. (2000). Conventional factor analysis vs. Rasch residual factor analysis *Rasch Measurement Transactions, 14*(2), 753.

Table 1: Number of papers available for each academic

Numbers of papers	**Number of academics**
1	2
2	1
3	1
4	12
5	10
6	8
>6	8
Total	**42**

Table 2: Number of ratings per paper

Numbers of times a paper was rated	Number of papers
1	21
2	56
3	53
4	58
5	29
6	5
9	1

Table 3: Number of levels recorded during the ratings process

Rating (rounded)	Number
0	21
1*	75
2*	294
3*	257
4*	63

Figure 1: Distribution of the ratings of papers

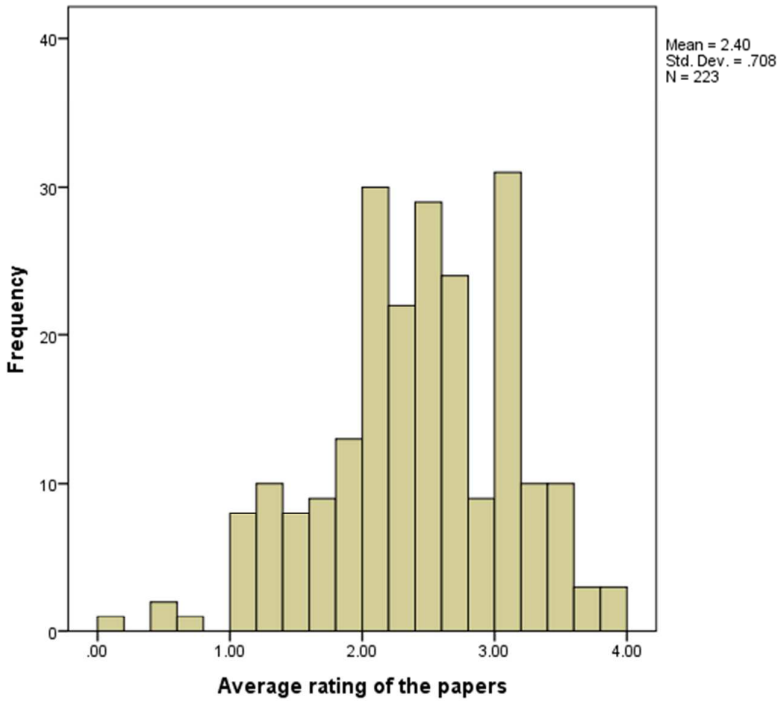

Figure 2: Distribution of the average ratings of the raters

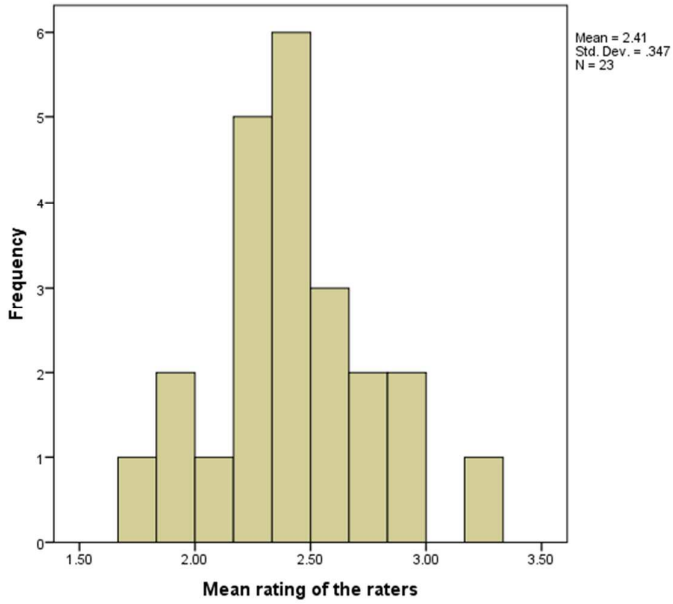

Figure 3: Probability of category probabilities half and full category ratings

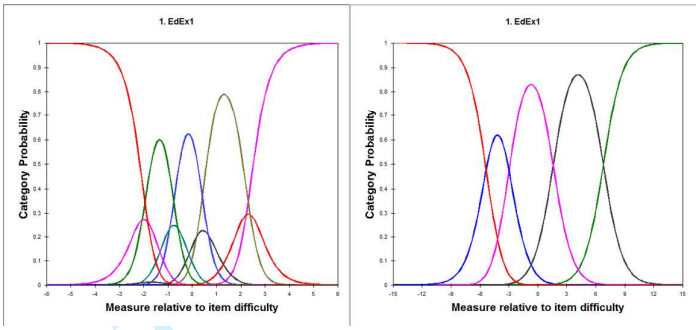

Figure4: Scattergram of Rasch measures of papers against the mean raw rating

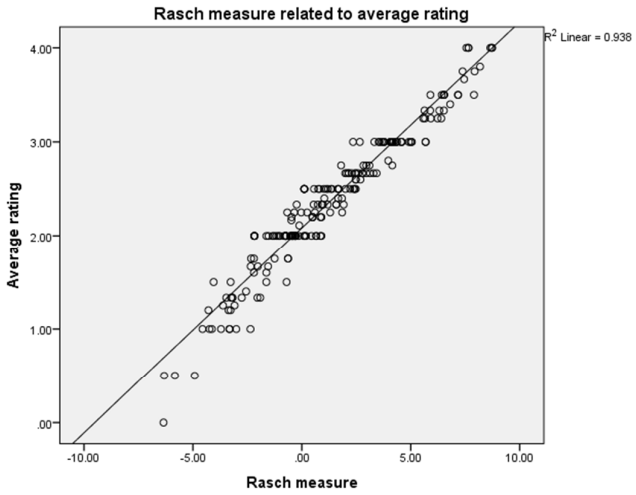

Figure 5: Error on the paper measures plotted against the level for different numbers of ratings

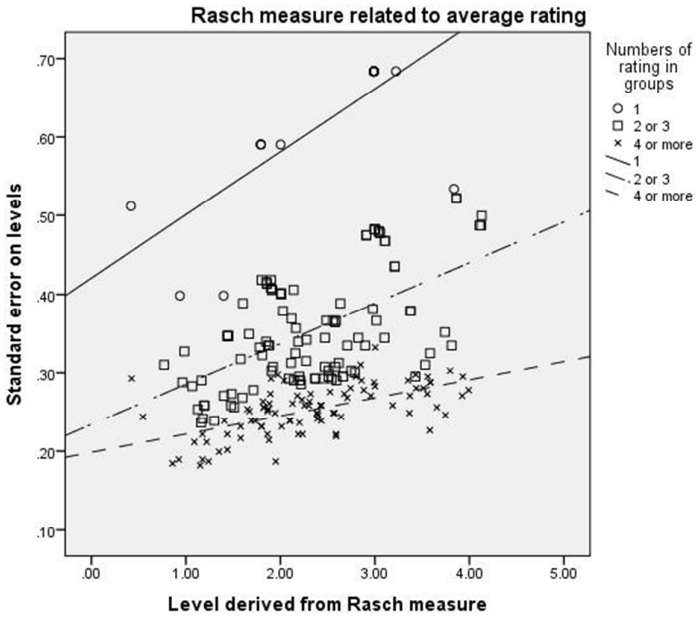

Figure 6: Map of paper level and rating severity

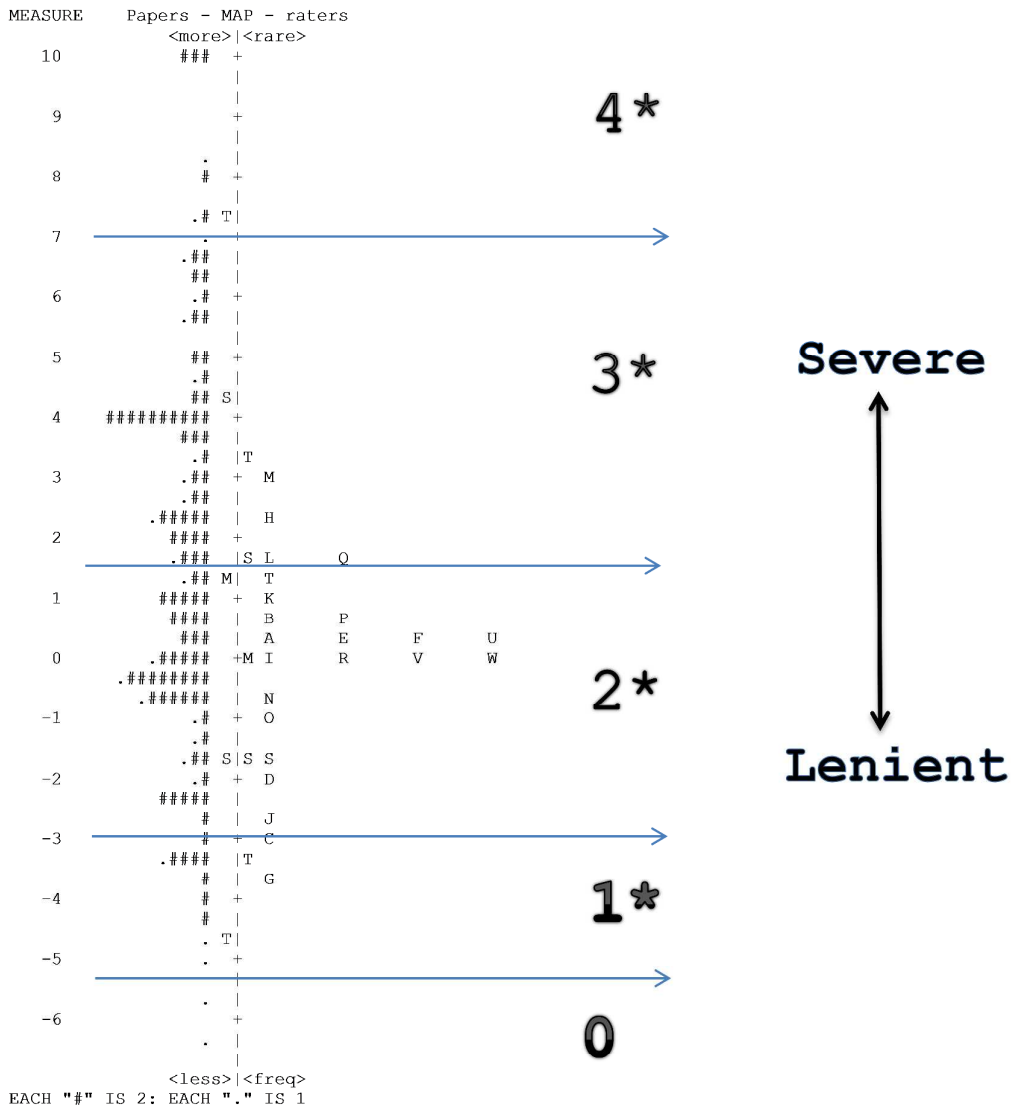

Figure 7 Rank order of the mean rating of each individual with errors of measurement